

#2

S&H Form: (2/01)

Attorney Docket No. 1086.1147

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of:

Hidesato MATSUOKA, et al.

Application No.: Unassigned

Group Art Unit: Unassigned

Filed: August 3, 2001

Examiner: Unassigned

For: DOCUMENT ANONYMITY SETTING DEVICE, METHOD AND COMPUTER
READABLE RECORDING MEDIUM RECORDING ANONYMITY SETTING PROGRAM



**SUBMISSION OF CERTIFIED COPY OF PRIOR FOREIGN
APPLICATION IN ACCORDANCE
WITH THE REQUIREMENTS OF 37 C.F.R. §1.55**

Assistant Commissioner for Patents
Washington, D.C. 20231

Sir:

In accordance with the provisions of 37 C.F.R. §1.55, the applicant(s) submit(s) herewith
a certified copy of the following foreign application:

Japanese Patent Application No. 2001-00641

Filed: January 5, 2001

It is respectfully requested that the applicant(s) be given the benefit of the foreign filing
date(s) as evidenced by the certified papers attached hereto, in accordance with the
requirements of 35 U.S.C. §119.

Respectfully submitted,
STAAS & HALSEY LLP

Date: August 3, 2001

By: _____

James D. Halsey, Jr.
Registration No. 22,729

700 11th Street, N.W., Ste. 500
Washington, D.C. 20001
(202) 434-1500

日 本 国 特 許 庁
JAPAN PATENT OFFICE



別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出 願 年 月 日

Date of Application:

2001年 1月 5日

出 願 番 号

Application Number:

特願2001-000641

出 願 人

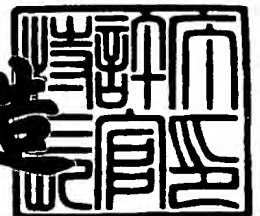
Applicant(s):

富士通株式会社

2001年 5月25日

特許庁長官
Commissioner,
Japan Patent Office

及 川 耕 造



出証番号 出証特2001-3045174

【書類名】 特許願

【整理番号】 0051814

【提出日】 平成13年 1月 5日

【あて先】 特許庁長官殿

【国際特許分類】 G11B 15/20

【発明の名称】 文書匿名化装置、方法及び匿名化プログラムを記録した
コンピュータ読取り可能な記録媒体

【請求項の数】 18

【発明者】

 【住所又は居所】 神奈川県川崎市中原区上小田中4丁目1番1号 富士
通株式会社内

 【氏名】 松岡 秀達

【発明者】

 【住所又は居所】 神奈川県川崎市中原区上小田中4丁目1番1号 富士
通株式会社内

 【氏名】 落谷 亮

【特許出願人】

 【識別番号】 000005223

 【氏名又は名称】 富士通株式会社

【代理人】

 【識別番号】 100079359

 【住所又は居所】 東京都港区西新橋3丁目25番47号 清水ビル8階

 【弁理士】

 【氏名又は名称】 竹内 進

 【電話番号】 03(3432)1007

【選任した代理人】

 【識別番号】 100093584

 【住所又は居所】 東京都港区西新橋3丁目25番47号 清水ビル8
階

【弁理士】

【氏名又は名称】 宮内 佐一郎

【電話番号】 03(3432)1007

【手数料の表示】

【予納台帳番号】 009287

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 9704823

【ブルーフの要否】 要

【書類名】 明細書

【発明の名称】 文書匿名化装置、方法及び匿名化プログラムを記録したコンピュータ読取り可能な記録媒体

【特許請求の範囲】

【請求項 1】

文書を入力する文書入力部と、

前記入力文書から匿名対象表記を抽出し、抽出した匿名対象表記がどの程度の強さで個人を特定できるかを評価する特定度を算出する特定度計算部と、

所定の閾値より大きい特定度を持つ前記入力文書中の表記を匿名化する匿名化処理部と、

を備えたことを特徴とする文書匿名化装置。

【請求項 2】

請求項 1 記載の文書匿名化装置に於いて、

前記特定度計算部は、前記入力文書から人名を抽出し、抽出した人名がどの程度の強さで個人を特定できるかを評価する特定度を算出し、

前記匿名化処理部は、所定の閾値よりも大きい特定度をもつ人名を匿名化することを特徴とする文書匿名化装置。

【請求項 3】

請求項 1 記載の文書匿名化装置に於いて、

前記特定度計算部は、前記入力文書から人名の周辺表記を抽出し、抽出された周辺表記がどの程度の強さで個人を特定できるかを評価する特定度を算出し、

前記匿名化処理部は、所定の閾値よりも大きい特定度をもつ周辺表記を匿名化することを特徴とする文書匿名化装置。

【請求項 4】

請求項 1 記載の文書匿名化装置に於いて、前記特定度算出部は、
入力文書から文を切出す文切出し部と、
切出した文を品詞毎に分解する品詞解析部と、
前記品詞解析結果から人名抽出ルールに基づいて人名を抽出する人名抽出部と
統計情報に基づいて抽出した人名の特定度を計算する人名特定度計算部と、
を備えたことを特徴とする文書匿名化装置。

【請求項 5】

請求項 1 記載の文書匿名化装置に於いて、前記特定度算出部は、更に、
前記品詞解析結果から構文解析ルールに基づいて文節間の係り受け関係を示す
構文木を作成する構文解析部と、
前記構文解析部で得られた構文木に対し個人特定木抽出ルールに基づいて個人
特定木を個人周辺表記として抽出する個人特定木抽出部と、
統計情報に基づいて抽出した個人特定木の特定度を計算する木構造特定度計算
部と、
を備えたことを特徴とする文書匿名化装置。

【請求項 6】

請求項 4 又は 5 記載の文書匿名化装置に於いて、前記特定度算出部は、既存文
書に基づいて作成した匿名対象表記、人名か周辺表記かの種別及び特定度を組に
した特定度データを登録した基準特定度データベースを備え、前記入力文書から
抽出した匿名化表記の計算により求めた特定度を、前記基準匿名度データベー
スに登録している特定度との重み平均をとって正規化することを特徴とする文書匿
名化装置。

【請求項 7】

請求項 4 又は 5 記載の文書匿名化装置に於いて、前記特定度算出部は、文書デ
ータベースの既存文書から文書毎に人名や周辺表記を抽出して特定度を計算し、
匿名対象表記、人名又は周辺表記の種類及び特定度の組にした特定度データを登

録した前記基準特定度データベースを作成するデータベース作成部を備えたことをことを特徴とする文書匿名化装置。

【請求項 8】

請求項 1 記載の文書匿名化装置に於いて、更に、前記匿名化処理部で使用する閾値を設定変更する匿名化指示部を設けたことを特徴とする文書匿名化装置。

【請求項 9】

請求項 1 記載の文書匿名化装置に於いて、前記匿名化指示部は、処理文書毎に匿名化処理に使用した閾値を閾値データベースに保存し、新たな入力文書の匿名化処理の際に直前の閾値をデフォルトとして設定することを特徴とする文書匿名化装置。

【請求項 1 0】

請求項 1 記載の文書匿名化装置に於いて、前記匿名化処理部は、匿名化不要表記を登録して匿名化不要データベースを持ち、入力文書から抽出された匿名化表記の内、前記匿名化不要データベースに登録されている表記は匿名化しないことを特徴とする文書匿名化装置。

【請求項 1 1】

請求項 1 記載の文書匿名化装置に於いて、前記匿名化処理部は、必ず匿名化する表記を登録した匿名化データベースを持ち、該匿名化データベースに登録されている入力文書中の表記は全て匿名化することを特徴とする文書匿名化装置。

【請求項 1 2】

請求項 1 記載の文書匿名化装置に於いて、前記匿名化処理部は、入力文書から抽出された匿名化対象表記を伏せ字にすることを特徴とする文書匿名化装置。

【請求項 1 3】

請求項 1 記載の文書匿名化装置に於いて、前記匿名化処理部は、入力文書から抽出された匿名化対象表記を、個人を特定しない一般化された表記に置き換えることを特徴とする文書匿名化装置。

【請求項 1 4】

請求項 1 記載の文書匿名化装置に於いて、前記匿名化処理部は、入力文書から抽出された匿名化対象表記を、該匿名化対象表記の匿名化に使用する閾値以下の特定度を持つ低特定度表記で置き換えることを特徴とする文書匿名化装置。

【請求項 1 5】

請求項 1 記載の文書匿名化装置に於いて、前記匿名化処理部は、入力文書から抽出された匿名化対象表記を暗号化することで匿名化することを特徴とする文書匿名化装置。

【請求項 1 6】

請求項 1 6 記載の文書匿名化装置に於いて、更に、前記匿名化処理部により暗号化により匿名化された匿名化文書を閲覧する際に、暗号化された匿名化表記を復号化して表示させる復号化指示部を設けたことを特徴とする文書匿名化装置。

【請求項 1 7】

文書を入力する文書入力ステップと、

前記入力文書から匿名対象表記を抽出し、抽出した匿名対象表記がどの程度の強さで個人を特定かを評価する特定度を算出する特定度計算ステップと、

所定の閾値より大きい特定度を持つ前記入力文書中の表記を匿名化する匿名化処理ステップと、

を備えたことを特徴とする文書匿名化方法。

【請求項 1 8】

コンピュータに、

文書を入力する文書入力ステップと、

前記入力文書から匿名対象表記を抽出し、抽出した匿名対象表記がどの程度の強さで個人を特定かを評価する特定度を算出する特定度計算ステップと、

所定の閾値より大きい特定度を持つ前記入力文書中の表記を匿名化する匿名化処理ステップと、

を実行させるための匿名化プログラムを記録したコンピュータ読取り可能な記録媒体。

【発明の詳細な説明】

【 0 0 0 1 】

【発明の属する技術分野】

本発明は、文書内の個人を特定するような表現を匿名化する文書匿名化装置、方法及び匿名化プログラムを記録したコンピュータ読取り可能な記録媒体に関し、特に、個人を特定する表現がどの程度の強さで個人を特定できるかを評価して匿名化する文書匿名化装置、方法及び匿名化プログラムを記録したコンピュータ読取り可能な記録媒体に関する。

【 0 0 0 2 】

【従来の技術】

近年、コンピュータを利用したデータ解析の傾向として、顧客からのアンケート回答、苦情、電子メール等の電子化された文書データから事業に役立つ情報を抽出しようとする機運が高まっている。しかし、これらの文書データには個人情報が含まれていることが多く、取扱を間違えると企業の存立に関わる問題となり得る。そこで、文書データを解析する前に、個人情報に関わる情報を適切に隠蔽することが必要となる。

【 0 0 0 3 】

従来、文書データ等に含まれる個人情報は人手により隠蔽化するか、あるいは機械処理が可能な個人名等の直接に個人を特定する表現を隠蔽化する等が行われている。

【 0 0 0 4 】

【発明が解決しようとする課題】

しかしながら、このような従来の個人情報の隠蔽化にあっては、記述されている個人名や個人に関連する周辺表記が、個人情報として保護される情報に属するものか、公的な人物に関する情報のように保護の必要がない情報なのかの区別が作業者にとって判別しづらいため、作業者によって個人情報隠蔽化の適切さが変化するという問題がある。

【 0 0 0 5 】

このため個人情報の隠蔽化を行う作業者の技能と知識は、ある水準を越えている必要があるため、人手による個人情報の隠蔽化は高コストになりやすい。

【 0 0 0 6 】

本発明は、個人情報の隠蔽化を機械化して作業コストを低減し、更に、必要に応じて隠蔽化の度合を調整可能とする文書匿名化装置、方法及び匿名化プログラムを記録したコンピュータ読取り可能な記録媒体を提供することを目的とする。

【 0 0 0 7 】

【課題を解決するための手段】

図 1 は本発明の概略説明図である。本発明は、文書匿名化装置であり、図 1 (A) のように、文書を入力する文書入力部 1 0 と、入力文書から匿名対象表記を抽出し、抽出した匿名対象表記がどの程度の強さで個人を特定できるかを評価する特定度を算出する特定度計算部 1 2 と、所定の閾値より大きい特定度を持つ入力文書中の表記を匿名化する匿名化処理部 1 8 とを備えたことを特徴とする。

【 0 0 0 8 】

このため本発明は、文書中の個人を特定するような表現に対して、それがどの程度の強さで個人を特定できるのかを匿名化を行う前に評価しておき、要求される匿名化の水準（閾値）に応じて情報を隠蔽化する。この結果、文書を必要な度合いで自動ないし半自動で匿名化でき、匿名化作業を効率化し作業コストを下げることができる。

【0009】

ここで特定度計算部18は、入力文書から人名を抽出し、抽出した人名がどの程度の強さで個人を特定できるかを評価する特定度を算出し、匿名化処理部18は、所定の閾値よりも大きい特定度をもつ人名を匿名化する。

【0010】

また特定度計算部12は、入力文書から人名の周辺表記抽出し、抽出された周辺表記がどの程度の強さで個人を特定できるかを評価する特定度を算出し、このとき匿名化処理部16は、所定の閾値よりも大きい特定度をもつ周辺表記を匿名化することを特徴とする。ここで周辺表記とは、例えば「大手A社の社長」のように個人名を強く示唆する表記のことである。

【0011】

特定度算出部12は、例えば入力文書から文を切出す文切出し部と、切出した文を品詞毎に分解する品詞解析部と、品詞解析結果から人名抽出ルールに基づいて人名を抽出する人名抽出部と、統計情報に基づいて抽出した人名の特定度を計算する人名特定度計算部とを備える。

【0012】

更に、特定度算出部12は、品詞解析結果から構文解析ルールに基づいて文節間の係り受け関係を示す構文木を作成する構文解析部と、構文解析部で得られた構文木に対し個人特定木抽出ルールに基づいて個人特定木を周辺表記として抽出する個人特定木抽出部と、統計情報に基づいて抽出した個人特定木（周辺表記）の特定度を計算する木構造特定度計算部とを備える。

【0013】

特定度算出部12は、例えば入力文書から文を切出す文切出し部と、切出した文を品詞毎に分解する品詞分解部と、品詞分解部から人名抽出ルールに基づいて人名を抽出する人名抽出部と、品詞解析部から構文解析ルールに基づいて人名に係る文節間の係り受け関係を示す構文木を作成する構文木解析部と、構文解析結果で得られた構文木に対し、個人特定木抽出ルールに基づいて構文特定木を周辺表記として取り出す個人特定木抽出部と、基準文書内での周辺表記と人名の組み合わせを持つ統計情報から、以下の方法で人名や周辺表記が個人を特定する度合

である特定度を計算する特定度計算部を備える。

【 0 0 1 4 】

ここで、周辺表記とは、構文解析の結果、人名と係り受け関係を持つ構文木のことであり、例えば「大手 A 社の〇〇社長」には〇〇社長という人名があって、それに係る修飾句である「大手 A 社の」が周辺表記である。

【 0 0 1 5 】

特定度算出部 1 2 は、匿名化対象文書に含まれる人名や周辺表記の組み合わせに対して、基準特定度データベース 1 4 中の人名や周辺表記が特定の個人を指す確率を読み出して、基準特定度データベース 1 4 内の全ての個人識別 I D について匿名化対象文書中の人名や周辺表記の組み合わせが持つ個人情報をも特定する強さである特定度の計算を行う。基準特定度データベース 1 4 には、個人識別する I D と共に人名や周辺表記がその個人を指す確率が登録されている。

【 0 0 1 6 】

個人識別 I D が p である表記数の最大を N としたとき、匿名化対象の文書中の人名や周辺表記の組み合わせが p を特定する度合である特定度 $K(p)$ の計算は、ここでは次式で行う。

【 0 0 1 7 】

$$K(p) = (\text{入力文書中の人名又は周辺表記が } p \text{ を指す確率の総和}) / N \quad \dots (1)$$

ただし、特定度 $K(p)$ の計算方法は、これに限定されるものではなく、入力文書の人名や周辺表記と一致しない人名や周辺表記を持つ個人識別 I D については特定度が低くなり、一致する人名や周辺表記が多い程、表記が表わす個人識別 I D の特定度が高くなる性質を持つ計算方法であればよい。

【 0 0 1 8 】

基準特定度データベース中で全ての p について特定度 $K(p)$ を計算しているので、特定度 $K(p)$ を最大にする p を求めることができ、入力文書や人名や周辺表記の組み合わせはその p を越えている可能性が最も高いことになり、最大の特定度 $K(p)$ がある基準値を越えている場合に、入力文書の人名や周辺表記に対し隠蔽化を行うことになる。

【0019】

ここで特定度の計算に使用する基準特定度データベース14の作成方法を説明する。基準特定度データベース14の作成は、データベース作成部15により行われる。データベース作成部15は、既存文書の集合である文書データベース72から文書切出し部によって文書を切り出し、次に文切り出し部によって文単位に分解し、品詞解析部、人名処理部および周辺表記処理部で人名や周辺表記を抽出し、それらがある個人を指す確率を計算し、基準特定度データベース14に、表記が指す個人を識別するID、表記の種類、表記、表記が個人を指す確率の4つの組でなる基準特定度データを登録する。

【0020】

確率計算のためには表記が指す個人を特定する必要があり、そのため電子メールアドレスや住所といった個人を特定する表記を使用する。これらは、以下ののような表記の特徴を持っており、その表記の特徴を利用して文書中から取り出す。

【0021】

(1) 電子メールアドレス : abcd@xxx.yyyy.com

(2) 住所 : ○○県○○市○○△丁目△△-△△

これらの個人を特定する表記を個人識別IDに変換する。ある個人識別IDであるpを使って、ある特定の個人を指す確率 $P(a \rightarrow p)$ や、人名の周辺表記sが、特定の個人を指す確率 $P(s \rightarrow p)$ を以下の式で近似する。文書データベース中で個人を特定する表記を持ち、それがpを指している文書の集合をMとすると、

$$P(a \rightarrow p) = (M \text{ におけるある } a \text{ の個数}) / (M \text{ における全ての } a \text{ の数}) \quad \dots (2)$$

$$P(s \rightarrow p) = (M \text{ におけるある } a \text{ の個数}) / (M \text{ における全ての } a \text{ の数}) \quad \dots (3)$$

となる。近似の方法は常識に限定されるものではない。

【0022】

この計算結果から、個人識別ID、表記の種類、表記、表記が個人を指す確率の4つの組のデータが基準特定度データベースに基準特定度データとして登録さ

れる。

【 0 0 2 3 】

基準特定度データベース 1 4 から個人を特定する確率を読み出して、特定度計算部で特定度を計算する。計算された特定度を基準値と比較することで、匿名化処理を行うかどうか判別する。

【 0 0 2 4 】

特定度算出部 1 2 は、文書データベースの既存文書から文書毎に人名や周辺表記を抽出して特定度を計算し、匿名対象表記、人名又は周辺表記の種類及び特定度の組にした特定度データを登録した基準特定度データベース 1 4 を作成するデータベース作成部 1 5 を備えるようにしても良い。

【 0 0 2 5 】

本発明の匿名化装置は、更に、匿名化処理部 1 8 で使用する閾値を設定変更する匿名化指示部 2 0 を設ける。このため作業者は、閾値を設定変更しながら匿名化された文書をチェックすることで、個人情報に対する隠蔽化度合を簡単に調整でき、最適な隠蔽化ができる。

【 0 0 2 6 】

匿名化指示部 2 0 は、処理文書毎に匿名化処理に使用した閾値データベース 2 6 に保存し、新たな入力文書の匿名化処理の際に直前の閾値をデフォルトとして設定する。このため検属して文書の匿名化処理をおこなう場合には、一度、最適な閾値の設定調整が済めば、その後は最適化された閾値がデフォルト設定さることで、特に閾値を指示する必要なく処理を進めることができる。

【 0 0 2 7 】

匿名化処理部 1 8 は、匿名化不要表記を登録した匿名化不要データベース 2 2 を持ち、入力文書から抽出された匿名化表記の内、匿名化不要データベースに登録されている表記は匿名化しない。例えば首相や大臣のように公的な人物については、匿名化データベースに登録することで、匿名化の対象から除外する。

【 0 0 2 8 】

匿名化処理部 1 8 は、必ず匿名化する表記を登録した匿名化データベース 2 4 を持ち、この匿名化データベース 2 6 に登録されている入力文書中の表記は全て

匿名化する。例えば企業名、クレジットカードの番号を表す規則、電話番号を表す規則、電子メールのアドレスを表す規則等については、匿名化データベース 26 に登録しておくことで、閾値の如何に関わらず確実に匿名化する。

【0029】

匿名化処理部 18 は匿名化処理として次の処理を選択的に行うことができる。

(1) 入力文書から抽出された匿名化対象表記を伏せ字にする。

(2) 入力文書から抽出された匿名化対象表記を、個人を特定しない一般化された表記に置き換える。

(3) 入力文書から抽出された匿名化対象表記を、匿名化対象表記の匿名化に使用する閾値以下の特定度を持つ低特定度表記で置き換える。

(4) 入力文書から抽出された匿名化対象表記を暗号化することで匿名化する。こごて、匿名化処理部 18 により暗号化により匿名化された匿名化文書を閲覧する際に、暗号化された匿名化表記を復号化して表示させる復号化指示部 32 を設けるようにしても良い。

【0030】

また本発明は、文書匿名化方法を提供するものであり、図 1 (B) のように、文書を入力する文書入力ステップと；

入力文書から匿名対象表記を抽出し、抽出した匿名対象表記がどの程度の強さで個人を特定できるかを評価する特定度を算出する特定度計算ステップと；

所定の閾値より大きい特定度を持つ入力文書中の表記を匿名化する匿名化処理ステップと；

を備えたことを特徴とする。この文書匿名化方法の詳細は装置構成の場合と同じになる。

【0031】

更に本発明は、匿名化プログラムを記録したコンピュータ読取り可能な記録媒体を提供するものであり、記録媒体に記録された匿名化プログラムは、コンピュータに、

文書を入力する文書入力ステップと、

入力文書から匿名対象表記を抽出し、抽出した匿名対象表記がどの程度の強さで

個人を特定できるかを評価する特定度を算出する特定度計算ステップと、
所定の閾値より大きい特定度を持つ入力文書中の表記を匿名化する匿名化処理ステップと、
を実行させる。この記録媒体における匿名化プログラムの詳細も装置構成の場合と同じになる。

【 0 0 3 2 】

【発明の実施の形態】

図 2 は、本発明による文書匿名化装置の機能構成を示したブロック図であり、コンピュータ装置のプログラム制御により実現される。

【 0 0 3 3 】

図 2 において、本発明の文書匿名化装置は、文書入力部 1 0、特定度計算部 1 2、基準特定度データベース 1 4、特定度正規化部 1 6、匿名化処理部 1 8、匿名化指示部 2 0、匿名化不要データベース 2 2、匿名化データベース 2 4、閾値データベース 2 6、作業表示部 2 8 及び匿名化文書記憶部 3 0 で構成される。更に匿名化文書記憶部 3 0 に格納された匿名化文書を閲覧するため、復号化指示部 3 2、判定部 3 4、閲覧データ作成部 3 6 及び閲覧表示部 3 8 が必要に応じて設けられる。

【 0 0 3 4 】

このような機能構成を持つ本発明の文書匿名化装置につき、各処理部の詳細を説明すると次のようになる。文書入力部 1 0 は匿名化対象文書を入力する。匿名化文書としては、例えばデータ解析対象となる文書が含まれ、例えば顧客からのアンケート回答、苦情、電子メールなどの文書情報を含んでいる。

【 0 0 3 5 】

文書入力部 1 0 で入力した匿名化対象文書は特定度計算部 1 2 に与えられ、特定度が計算される。ここで特定度とは、個人を特定するような表現、即ち人名やその周辺表記に対して、どの程度の強さで個人を特定できるかを評価する値である。

【 0 0 3 6 】

本発明にあっては、特定度算出部 12 は、匿名化対象文書から人名や周辺表記を抽出し、抽出した人名や周辺表記が個人を指す確率を文書データベース 72 から前記 (2) 式や (3) 式に基づいて算出するか、もしくは、基準特定度データベースから表記が個人を指し示す確率を読み出して特定度を計算し、匿名化状態部 18 は、所定の閾値よりも大きい特定度を持つ神明や周辺表記を隠蔽化する。

【0037】

基準特定度データベース 14 には、データベース作成部 15 により、後の説明で明らかにするように、十分な量の文書データベース 72 を使用して、そこに存在している人名や周辺表記について、前記 (2) 式や (3) 式から表記が個人を指す確率を算出して登録している。

【0038】

ここで基準特定度データベース 14 からは、図 9 に示すような

- (1) 個人識別 ID
- (2) 個人識別 ID を示す表記の種別 (人名もしくは周辺表記)
- (3) 個人識別 ID を指す表記
- (4) 表記が個人識別 ID を指す確率

の 4 つを組としたデータが出力される。特定度計算部 12 は、基準特定度データベース 14 の出力から、(1) 式に基づいてある個人識別 ID について特定度を算出する。

【0039】

図 9 に示した基準特定度データベース 14 の例に従って説明する。この例では前記 (1) 式で使用する N は 4 であったとする。そして P001 の個人を特定する度合である特定度は、(1) 式から

$$(0. + 0.9 + 1.0 + 0.2) / 4 = 0.6$$

となる。P003 については、人名「松岡」のみが一致したとすると、(1) 式から $0.2 / 4 = 0.05$ が P003 についての特定度となる。全ての個人識別 ID について計算した特定度の中で P001 の 0.6 が最大だったとすると、この 0.6 と基準値を比較して隠蔽化するかどうかの判定を行う。

【0040】

特定度算出部 1 2 は、

- (1) 匿名化表記
- (2) 特定度

2 つの組となる特定度データの形式で匿名化処理部 1 8 に出力する。

【 0 0 4 1 】

匿名化処理部 1 8 は特定度正規化部 1 6 より出力された特定度データを使用して、文書入力部 1 0 より得られた匿名化対象文書について人名や周辺表記を隠蔽化する匿名化処理を行う。

【 0 0 4 2 】

匿名化処理部 1 8 に対しては、匿名化指示部 2 0 より

- (1) 閾値
- (2) 使用匿名化方法
- (3) 処理文書分類

の 3 つの指示値が与えられる。

【 0 0 4 3 】

匿名化指示部 2 0 による閾値は、作業者がキーボードやマウスなどの入力デバイスを使用して設定変更できる値であり、人名閾値と周辺表記を対象とした個人特定木閾値を個別に設定することができる。この匿名化指示部 2 0 から設定された閾値は特定度正規化部 1 6 から得られた特定度データの特定度と比較され、閾値以上の特定度をもつ人名及び周辺表記について隠蔽するための匿名化処理が行われる。

【 0 0 4 4 】

匿名化指示部 2 0 による閾値の設定は、直接閾値の数値を入力させる方法以外に、特定度に対応したスライダーをマウスで操作する方法、閾値の設定ウィンドウを開いてウィンドウ項目の中から閾値を選択する方法など、適宜の視覚的な操作を含む。

【 0 0 4 5 】

匿名化指示部 2 0 による使用匿名化方法の指示に対応して、匿名化処理部 1 8 には次の匿名化方法が設けられている。

- (1) 伏せ字化
- (2) 一般化
- (3) 低特定度化
- (4) 暗号化

まず伏せ字化は、匿名化対象とする人名や周辺表記を伏せ字に使用する記号を選択し、選択した記号で匿名化対象を全て置き換える。例えば「佐藤」といった人名を「××」とする。

【0046】

一般化は匿名化対象中の固有名詞を一般的な表記で置き換える。このため、一般的な表記で置き換えるための一般化ルールを匿名化処理部18は備えている。この一般化ルールには例えば

ルール1：人名は「A」に置き換える。

【0047】

ルール2：企業名は「A」に置き換える。

などが記述されている。

【0048】

低特定度化は、匿名化の対象をより特定度の低い表記で置き換える。この低特定度化のため、人名と特定度の組及び個人特定木と特定度の組について、小さい特定度のものを基準特定度データベース14から検索し、このとき匿名化対象となっている特定度より低い特定度の検索した表記を用いて匿名化対象を置き換える。具体的には、人名の場合には基準特定度データベース14から小さい特定度を持つ人名を検索し、検索結果で匿名化対象の人名を置き換えることで匿名化する。

【0049】

また周辺表記となる個人特定木の場合には、基準特定度データベース14から匿名化対象の個人特定木と共通する分節を含む個人特定木で特定木が小さい個人特定木を検索し、この検索した個人特定木で匿名化対象を置き換える。もし基準特定度データベース14から検索した結果が匿名化対象の特定度以下でない場合には、低特定度化による匿名化はできないことから、処理の失敗を作業者に知ら

せることになる。更に、暗号化は、匿名化の対象を所定の暗号規則に従って暗号化する。

【 0 0 5 0 】

匿名化処理部 1 8 に対しては、匿名化不要データベース 2 2 と匿名化データベース 2 4 が設けられている。匿名化不要データベース 2 2 には匿名化を行う表記や識別するための規則が登録されている。匿名化不要データベース 2 2 に登録されている例としては例えば次のものがある。

- (1) 首相、大臣などの公人の人名
- (2) 芸能人の人名
- (3) 首相、大臣のような、公の人物であることを示す周辺表記を持つ人物を識別する規則

このため、匿名化処理部 1 8 で文書入力部 1 0 より入力した匿名化対象文書を匿名化する際に、匿名化不要データベース 2 2 を参照し、そこに登録している人名や表記については匿名化を一切行わないことになる。

【 0 0 5 1 】

匿名化データベース 2 4 には匿名化処理の際に必ず匿名化を行う表記や、これを識別するための規則が登録されている。例えば匿名化データベース 2 4 には次の表記や規則が登録されている。

- (1) 企業名
- (2) クレジットカードの番号を表わす規則
- (3) 電話番号を表す規則
- (4) 電子メールのアドレスを表わす規則

このため匿名化処理部 1 8 にあっては、文書入力部 1 0 から入力した匿名化対象文書の中に匿名化データベース 2 4 に登録している表記や識別規則に該当する表記がある場合には、特定度計算部 1 2 及び特定度正規化部 1 6 で求められた特定度の如何に関わらず強制的に匿名化を行って認定することになる。

【 0 0 5 2 】

匿名化処理部 1 8 で匿名化された文書は作業表示部 2 8 に表示され、作業者は匿名化の結果を確認しながら匿名化指示部 2 0 により閾値や使用匿名化方法を変

更し、必要とする隠蔽化が行われた匿名化文書を作成することができる。匿名化処理部 1 8 で作成された匿名化文書は匿名化文書記憶部 3 0 に保存される。匿名化文書記憶部 3 0 のレコード形式としては、文書コードなどを使用した処理文書分類、匿名化処理情報、匿名化文書の形式をもって保存する。もちろん指示情報には、文書匿名化処理の際に匿名化指示部 2 4 で指示された閾値使用匿名化方法が含まれている。

【 0 0 5 3 】

閾値データベース 2 6 には匿名化文書記憶部 3 0 に記憶された匿名化文書レコードより得た閾値と匿名化方法が処理文書分類である分類コードによって登録されている。特に閾値データベース 2 6 の先頭位置には、最新の匿名化文書に関する閾値により匿名化方法が格納されており、匿名化処理部 1 8 にあっては、この閾値データベース 2 6 の先頭位置の閾値及び匿名化方法を匿名化指示部 2 0 によるデフォルトの設定内容としている。

【 0 0 5 4 】

このため作業者にあっては、匿名化指示部 2 0 により閾値や使用匿名化方法の指示を行わなくても、直前に行われた匿名化文書における閾値及び匿名化方法が自動的に匿名化処理部 1 8 に設定されることになる。

【 0 0 5 5 】

匿名化文書記憶部 3 0 に保存されている匿名化文書は、閲覧データ作成部 3 6 により読み出して閲覧表示部 3 8 に表示して閲覧することができる。このうち暗号化による匿名化文書については、復号化指示部 3 2 からの暗号化方法に対応したパスワードの入力で匿名化文書中の暗号化部分を元の人名や周辺表記に復号化して閲覧することができる。

【 0 0 5 6 】

復号化指示部 3 2 からのパスワードは判定部 3 4 で判定され、パスワードに対応した復号化方法が閲覧データ作成部 3 6 に指示され、暗号化された表記を復号して閲覧することができる。この匿名化処理部 1 8 における暗号化と閲覧時の復号化については、後の説明で更に明らかにされる。

【 0 0 5 7 】

図 3 は、図 2 における本発明の文書匿名化処理のフローチャートである。図 3 において、ステップ S 1 で匿名化処理要求の有無をチェックしており、作業者による匿名化処理要求を判別すると、ステップ S 2 に進み、文書入力部 1 0 より匿名化対象文書を入力する。続いてステップ S 3 で特定度計算部 1 2 により匿名化対象文書に含まれる人名やその周辺表記である匿名化対象について特定度を計算する。具体的には、基準特定度データベース 1 4 の参照で個人名の指す確率を取得する。続いてステップ S 4 で特定度に基づいて匿名化処理を行う。

【 0 0 5 8 】

ステップ S 4 で匿名化処理が済むと、ステップ S 5 で匿名化文書を保存する。続いてステップ S 6 で閲覧要求をチェックしており、閲覧要求があればステップ S 7 に進み、保存している匿名化文書の閲覧データを作成して表示する。そしてステップ S 8 で終了指示があれば、一連の処理を終了する。

【 0 0 5 9 】

図 4 は、図 2 の特定度計算部 1 2 の詳細を示した機能構成のブロック図である。特定度計算部 1 2 は、文切出部 4 0、品詞解析部 4 2、人名処理部 4 4 及び周辺表記処理部 4 6 で構成される。人名処理部 4 4 には人名抽出部 4 8、人名特定度計算部 5 0 及び人名抽出ルール 5 2 が設けられている。

【 0 0 6 0 】

また周辺表記処理部 4 6 には構文解析部 5 4、個人特定木抽出部 5 6、木構造特定度計算部 5 8、構文解析ルール 6 0 及び個人特定木抽出ルール 6 2 が設けられている。

【 0 0 6 1 】

図 2 の文書入力部 1 0 で入力された匿名化文書は、図 4 の特定度計算部 1 2 における文切出部 4 0 に与えられ、文単位に分解して切り出した文を品詞解析部 4 2 に入力する。品詞解析部 4 2 は形態素解析などを利用して切り出した文を品詞情報付きの品詞に分解し、人名処理部 4 4 と周辺表記処理部 4 6 のそれぞれに出力して人名処理及び周辺表記処理をそれぞれ独立に行わせる。

【 0 0 6 2 】

まず人名処理部 4 4 を説明すると、品詞解析部 4 2 から文を品詞ごとに分解し

て受けた人名抽出部48は、人名抽出ルール52を用いて人名を抽出し、人名特定度計算部50に出力する。人名抽出ルール52としては「if～then～」形式によって次の規則が登録されている。

規則521： if [姓] , [名] then 人名として抽出

規則522： if [姓] then 人名として抽出

規則523： if [名] then 人名として抽出

この「if～then」の規則において、ifの次の条件部では品詞名を[]で表わす。また、この条件部で「,」で繋がった品詞は連続しているものを表わしている。

【0063】

このような人名抽出ルール60の規則521, 522, 523によって、規則に一致する品詞パターンを持った文字列として人名が抽出される。例えば規則521により連続した姓名から人名が抽出される。また規則522により姓から人名が抽出される。更に規則523により名から人名が抽出される。

【0064】

人名特定度計算部50は、匿名化対象となる表記により、基準特定度データベース14を参照し、表記が個人を指す確率を取得する。

【0065】

次に周辺表記処理部46に説明する。周辺表記処理部46の構文解析部56は、品詞解析部42から得られた品詞ごとに分解した文を対象に、構文解析を利用して分節間の係り受け関係を示す木構造、即ち構文木を作成する。このとき構文解析部54は構文解析ルール60を使用する。構文解析ルール60には「if～then」形式で次の規則が記述されている。

規則601： if [名詞句] , [助詞「の」] , [人名] then

[人名名詞句] ([名詞句] →<修飾>→ [人名])

規則602： if [名詞句] , [助詞「の」] , [人名名詞句] then

[人名名詞句] ([名詞句] →<修飾>→ [人名名詞句])

この規則601, 602において、ifの後ろの条件部は品詞間に複数の要素が入っている条件を表している。またthenの直後には、条件部が成立した場合

にひとまとめにした品詞を記述し、（ ）内に生成する要素間の関係を記述する。更に< >の中には生成される関係に付けられた名前を表わしている。

【0066】

この構文解析による構文木の生成を具体的に説明すると次のようになる。いま次のような文があったとする。

「△△社の社長でピアニストの〇〇は××ホールで演奏した」

規則 601 は名詞句と名詞句の間に助詞の「の」が入っているときに全体を名詞句とし、

[名詞句] →<修飾>→ [人名]

の修飾関係を生成し、これは例文の「ピアニストの」が「〇〇」を修飾している木構造に対応している。したがって、この場合の木構造として図 6 が得られる。

【0067】

図 6 のように得られた構文木に対し、次の個人特定木抽出部 56 は、個人特定木抽出ルール 62 を適用して、個人を特定する部分木を個人特定木として抽出し、木構造特定度計算部 58 に出力する。

【0068】

個人特定木抽出ルール 62 には次のような規則が登録されている。

規則 621 : if [名詞句] →<修飾>→<人名> then

個人特定木として抽出

規則 622 : if [名詞句] →<修飾>→ [人名名詞句] then

個人特定木として抽出

即ち規則 621 は、人名などを修飾する名詞句を個人特定木として抽出することに対応する。例えば「ピアニスト」が「〇〇」を修飾している木構造から

「ピアニストの」→<修飾>→「〇〇」

を個人特定木として抽出することができる。この例では、これ以外に図 7 のような木構造がそれぞれ個人特定木として抽出される。

【0069】

木構造特定度計算部 58 は人名特定度計算部 50 と同様、基準特定度データベース 14 の参照により、木構造が個人を指す確率を取得し、特定度を計算する。

【0070】

図7は、図4の特定度計算部12における処理のフローチャートである。この特定度計算処理にあっては、ステップS1で匿名化対象文書から文を切り出し、ステップS2で品詞ごとに分解する品詞解析を行い、人名処理及び周辺表記処理のそれぞれに供給する。

【0071】

人名処理にあっては、ステップS3で人名抽出を行い、ステップS4で人名特定度を計算し、併せて基準特定度データベース14の参照で得られた特定度と共に出力する。また周辺表記処理にあっては、ステップS5で構文解析を行った後、ステップS6で個人特定木抽出処理を行い、ステップS7で木構造特定度計算を行うと共に、基準特定度データベース14から基準特定度を取得し、正規化処理に出力する。

【0072】

図8は、図1の特定度計算部12に設けているデータベース作成部15の機能を取り出している。このデータベース作成部15は、文書データベース72に格納されている十分な量の文書を対象に基準特定度データベース14を作成する。

【0073】

このためデータベース作成部15にあっては、文書データベース72から対象文書を切り出す文切出部40が設けられ、切り出した文書は文切出部40に与えられる。

【0074】

データベース作成部15の文切出部40と品詞解析部42は、図4の特定度算出部12のブロックのものと同一のものが使用される。周辺表記処理部46-1は、個人を特定する周辺表記である電子メールアドレスや住所等を抽出し、それを個人識別IDに置き換える。

【0075】

電子メールアドレスや住所は、以下のような表記の特徴をっており、この表記を使って本分から取り出す。

【0076】

電子メールアドレス：abcd@xxx.yyyy.com

住所：〇〇県〇〇市〇〇△丁目△△-△△

人名処理部 4 4 - 1 及び周辺表記処理部 4 6 - 1 では、図 4 の人名処理部 4 4 及び周辺表記処理部 4 6 と同じ機構で人名や周辺表記を抽出する。抽出した人名や周辺表記については、前記（2）式や（3）式に従って、人名や周辺表記が個人を指す確率が計算される。

【 0 0 7 7 】

そして人名処理部 4 4 - 1 および周辺表記処理部 4 6 - 1 で作成された、個人識別 I D、表記の種類、表記、及び表記が個人を指す確率の 4 つの組となる特定度データは、例えば図 9 のように、基準特定度データベース 1 4 に基準特定度データとして登録される。

【 0 0 7 8 】

データベース作成部 1 5 の処理の流れを図 1 0 に示す。既存文書を厚めた文書データベース 7 2 から文書を切り出し、その文書を文に分解して人名と周辺表記である個人特定木を取り出すところまでは、図 5 の特定度計算部と同様である。周辺表記の中で個人を特定する電子メールアドレスや住所等の表記の特徴から判別して個人識別 I D を作成し、個人識別 I D 毎に、人名や周辺表記が個人を指す確率を（2）式もしくは（3）式から計算し、図 9 のような 4 つの組のデータとして基準特定度データベースに登録する。

【 0 0 7 9 】

図 9 は、基準特定度データベース 1 4 の登録内容の例であり、種別、表記、基準特定度の項目によって基準特定度データが登録され、種別としては人名及び周辺表記を表す構文木が格納されている。

【 0 0 8 0 】

図 1 0 は、図 8 のデータベース作成部 1 5 の処理のフローチャートである。この基準特定度データベースの作成処理にあっては、ステップ S 1 で文書データベース 7 2 から文書を切り出して標準文書を作成し、ステップ S 3 で品詞ごとに分解する品詞解析を行う。

【 0 0 8 1 】

この品詞解析結果はステップ S 4, S 5 の人名処理及びステップ S 6 ~ S 9 の周辺表記処理のそれぞれに与えられ、独立に人名抽出と人名特定度の計算、及び構文解析、個人特定木抽出に基づく木構造特定度、個人識別 ID の作成が行われる。そしてステップ S 1 0 で最終的に、基準特定度データベース 1 4 に図 9 のように基準特定度データを登録する。

【 0 0 8 2 】

この基準特定度データベースの作成処理は、本発明の匿名化装置を使用する前の準備段階で基本的に行うが、運用中においても必要に応じて適宜に文書データベース 7 2 を更新して、新たな文書データを対象に基準特定度データベース 1 4 の再構築を行うことが望ましい。

【 0 0 8 3 】

図 1 1 は、図 3 のステップ S 5 における匿名化処理の詳細のフローチャートである。この匿名化処理にあつては、ステップ S 1 で匿名化指示部 2 0 からの指示に基づき、匿名化処理部 1 8 で使用する匿名化情報を決定する。匿名化条件は匿名化指示部 2 0 からの指示がないときは、閾値データベース 2 6 に基づいて行う。

【 0 0 8 4 】

図 1 2 は、図 2 の閾値データベース 2 6 の登録内容であり、処理文書分類となる分類コード、閾値及び匿名化方法の項目で構成されている。この内、処理文書分類の分類コード 0 0 となる先頭位置には、直前の匿名化処理で使用した直前閾値とその匿名化方法が登録されている。この分類コード 0 0 の先頭位置の閾値及び匿名化方法は、図 1 1 の匿名化処理における匿名化指示のデフォルト条件として設定される。

【 0 0 8 5 】

このためステップ S 1 の匿名化条件決定の際に匿名化指示部 2 0 による作業者の指示がなければ、図 1 2 の閾値データベース 2 6 の先頭位置となる分類コード 0 0 の匿名化方法、この場合には「伏せ字化」と「直前閾値」が匿名化条件として設定される。

【 0 0 8 6 】

ステップ S 2 で匿名化条件が承認されると、ステップ S 3 で匿名化条件を決定した条件に変更する。そしてステップ S 4 で匿名化対象文書について匿名化表記である人名や周辺表記を検索し、ステップ S 5 で匿名化表記があれば、ステップ S 6 で匿名化不要データベース 2 2 の参照により匿名化不要表記を検索する。

【 0 0 8 7 】

ステップ S 7 で匿名化不要表記があれば、それ以降の処理をスキップする。匿名化不要表記がなければ、ステップ S 8 で匿名化表記について求められている特定度を匿名化条件として設定した閾値と比較し、閾値以上であればステップ S 9 の置換処理に入る。

【 0 0 8 8 】

この置換処理は、伏せ字化、一般化、低特定度化、暗号化のいずれかの処理となる。そしてステップ S 1 0 で全ての匿名化表記検索が終了したか否かチェックし、終了していなければ再びステップ S 4 に戻り、同様な処理を繰り返し、全ての匿名化表記の処理が済めば一連の処理を終了する。

【 0 0 8 9 】

図 1 3 は、図 1 1 のステップ S 9 における置換処理の詳細のフローチャートである。図 1 3 において、まずステップ S 1 で匿名化条件として伏せ字化の指示の有無をチェックしており、伏せ字化の指示であればステップ S 2 に進み、予め準備された伏せ字に使用する記号を選択し、ステップ S 3 で閾値以上の特定度を持つ匿名化対象表記を対象に伏せ字への置き換えを行う。

【 0 0 9 0 】

一方、ステップ S 4 で匿名化条件として一般化の指示が判別された場合には、予め準備している一般化ルールを参照して、ステップ S 5 で一般表記を選択し、ステップ S 6 で閾値以上の特定度を持っている匿名化対象表記について、選択した一般表記への置き換えを行う。

【 0 0 9 1 】

またステップ S 7 で匿名化条件として低匿名化の指示を判別した場合には、ステップ S 1 0 に進み、基準特定度データベース 1 4 より小さい特定度を持つ人名または周辺表記としての個人特定木をステップ S 1 0 で検索する。

【0092】

このデータベース検索に対し、ステップS11で低特定度表記があれば、ステップS12で検索した特定度表記への置き換えを行う。一方、ステップS11で低特定度表記がデータベースから検索できなかった場合には、ステップS14で作業者に対し失敗を通知して処理を終了する。

【0093】

一方、ステップS7で低特定度化でなかった場合には、この場合は暗号化であることからステップS8に進み、暗号化表記を生成し、ステップS9で匿名化対象表記を暗号化表記に置き換える。

【0094】

そしてステップS3、ステップS6、ステップS9またはステップS12のいずれかの置き換えが済むと、ステップS13で匿名化文書と表記データを出力し、必要があれば再度、匿名化条件の設定を行って匿名化処理を繰り返し、匿名化終了であれば匿名化文書記憶部30に匿名化文書を保存するようになる。

【0095】

ここでステップS8、S9における復号化による匿名化処理を説明すると次のようになる。暗号化表記による置き換えとしては、例えば暗号化によって匿名化した箇所の開始位置に暗号化したことを示すコード<CRYPT>を埋め込み、終了位置に暗号化の範囲が終了したことを示すコード</CRYPT>を埋め込む。また復号化方法を示す場合には開始コード<CRYPT>の部分を<CRYPT METHOD="復号化方法">として復号化方法を記述する。

【0096】

例えば「△△さんはプログラムの解析を行った」を暗号化により匿名化すると、次のようになる。

「<CRYPT METHOD="METHOD1">%abc\$12DE;KsrBX </CRYPT>さんはプログラムの解析を行った。」

この暗号化は匿名化対象文書の「△△」を暗号化した結果が

「%abc\$12DE;KsrBX」、復号化の方法が「METHOD1」の場合に、匿名化対象の表記「△△」を暗号化表記で置き換えたものである。

【0097】

この例では「METHOD1」で指定された復号化方法の中にパスワードや公開範囲を指定しておき、暗号化後のデータが外部に流出したとしても、匿名化対象「△△」という人名は復号化しない限り読み取ることができないようにする。

【0098】

また暗号化と復号化の方法を何通りか用意しておき、それぞれの方法ごとに対応するパスワードを変えておくことで、暗号化した表記ごとに復号化される部分とそうでない部分とを区別できるようにし、閲覧者ごとに読取り可能な範囲を変化させることもできる。更に暗号化の方法を記述する方法として、暗号化された匿名化文書に暗号化部分を示す情報を埋め込む方法以外に、暗号化した表記の位置情報や暗号化方法を匿名化文書、本文とは別文書で記憶させてもよい。

【0099】

このような暗号化による置換処理で得られた匿名化文書については、図2の復号化指示部32、判定部34、閲覧データ作成部36に示すように、暗号化方法と復号化方法に対応して定められたパスワードを使用した復号化指示部32からの指示を判定部34に対し行うことで、パスワードに基づいた復号化方法より匿名化文書記憶部30に格納されている暗号化表記で置換した匿名化文書を読み出し、暗号化表記の部分を元の人名や周辺表記に復号して閲覧表示部38で見ることができる。

【0100】

図14は、図2の作業表示部28に表示された匿名化作業画面88であり、文書入力部10より入力された匿名化対象文書として電子メール90が表示されている。この匿名化作業画面88の右側には匿名化条件を設定するウィンドウ92が設けられ、ウィンドウを開くことで原文92-1が表示されていることを示している。

【0101】

このような原文の匿名化作業画面88について、図15のようにウィンドウ92を開いて、その選択内容から閾値として低レベル92-2を指示し、この状態で実行キー94をマウスクリックすると、閾値レベルを低レベルとした本発明に

よる文書匿名化処理が実行され、匿名化文書 9 6 の表示が行われる。

【0 1 0 2】

この低レベルの閾値を設定した匿名化文書 9 6 を図 1 4 の原文である電子メール 9 0 と対比すると、企業名「情報媒体」、所属名「情報機器」、名「英達」が、それぞれ「〇〇〇〇」、「××××」「△△」に置換されている。また原文の電子メール 9 0 におけるメールアドレス、電話番号、ファックス番号及び住所についても、それぞれ匿名化の表記での置換が行われている。

【0 1 0 3】

図 1 6 は、ウィンドウ 9 2 の閾値レベルを高レベル 9 2 - 3 に設定した場合の実行キー 9 4 のマウスクリックによる処理結果としての匿名化文書 9 6 を表示した匿名化作業画面 8 8 である。

【0 1 0 4】

このように閾値レベルを高レベルとした場合には、図 1 5 の閾値レベルを低レベルとしたい場合には匿名化されていなかった人名「佐藤」「松岡」についても、「▽▽」「△△」のように匿名化表記への置換が行われ、個人情報に対する隠蔽度が更に高められる。

【0 1 0 5】

次に本発明による文書匿名化プログラムを記録したコンピュータ読取り可能な記録媒体の実施形態を説明する。本発明による文書匿名化プログラムは、図 3 のフローチャートに示した処理ステップを備えている。

【0 1 0 6】

即ち本発明の記録媒体に格納された匿名化プログラムは、コンピュータに文書を入力する文書入力ステップと、入力文書から匿名化対象表記を抽出し、抽出した匿名化対象表記がどの程度の強さで個人を特定するかを評価する特定度を参照する特定度計算ステップと、所定の閾値より大きい特定度を持つ入力文書中の表記を匿名化する匿名化処理ステップとを実行させる。

【0 1 0 7】

この記憶媒体には、CD-ROM やフロッピーディスクなどのリムーバブルな可搬型記録媒体、回線によりプログラムを提供するプログラム提供者の記憶装置

、更にはプログラムをインストールした処理装置のRAMやハードディスクなどのメモリ装置がある。また記録媒体によって提供された文書匿名化プログラムは処理装置にローディングされ、その主メモリ上で実行される。

【0108】

また本発明により提供される記録媒体に格納された文書匿名化プログラムは、図2における文書入力部10、特定度計算部12、基準特定度データベース14、特定度正規化部16、匿名化処理部18、匿名化指示部20、匿名化不要データベース22、匿名化データベース24及び閾値データベース26、更に匿名化文書記憶部30の処理機能を備えればよい。

【0109】

なお上記の実施形態にあつては、図2のように特定度計算部12で匿名化対象表記について特定度を計算と基準特定度データベース14の両方から求め、特定度正規化部16で正規化する場合を例にとっているが、特定度計算部12で匿名化表記について基準特定度データベース14から取得し、基準特定度データベース14にない場合に計算により特定度を求めるようにしてもよい。この場合には計算または基準特定度データベース14のいずれかから特定度が求まることから、特定度正規化部16による正規化は行わない。

【0110】

また本発明は、その目的と利点を損なわない適宜の変形を含む。更に本発明は上記の実施形態に示した数値による限定は受けない。

【0111】

(付記)

(付記1)

文書を入力する文書入力部と、
前記入力文書から匿名対象表記を抽出し、抽出した匿名対象表記がどの程度の強
さで個人を特定できるかを評価する特定度を算出する特定度計算部と、
所定の閾値より大きい特定度を持つ前記入力文書中の表記を匿名化する匿名化処
理部と、
を備えたことを特徴とする文書匿名化装置。(1)

(付記 2)

付記 1 記載の文書匿名化装置に於いて、

前記特定度計算部は、前記入力文書から人名を抽出し、抽出した人名がどの程度の強さで個人を特定できるかを評価する特定度を算出し、

前記匿名化処理部は、所定の閾値よりも大きい特定度をもつ人名を匿名化することを特徴とする文書匿名化装置。(2)

(付記 3)

付記 1 記載の文書匿名化装置に於いて、

前記特定度計算部は、前記入力文書から人名の周辺表記を抽出し、抽出された周辺表記がどの程度の強さで個人を特定できるかを評価する特定度を算出し、

前記匿名化処理部は、所定の閾値よりも大きい特定度をもつ周辺表記を匿名化することを特徴とする文書匿名化装置。(3)

(付記 4)

付記 1 記載の文書匿名化装置に於いて、前記特定度算出部は、

入力文書から文を切出す文切出し部と、

切出した文を品詞毎に分解する品詞解析部と、

前記品詞解析結果から人名抽出ルールに基づいて人名を抽出する人名抽出部と、

統計情報に基づいて抽出した人名の特定度を計算する人名特定度計算部と、

を備えたことを特徴とする文書匿名化装置。(4)

(付記 5)

付記 1 記載の文書匿名化装置に於いて、前記特定度算出部は、更に、

前記品詞解析結果から構文解析ルールに基づいて文節間の係り受け関係を示す構文木を作成する構文解析部と、

前記構文解析部で得られた構文木に対し個人特定木抽出ルールに基づいて個人特定木を個人周辺表記として抽出する個人特定木抽出部と、

統計情報に基づいて抽出した個人特定木の特定度を計算する木構造特定度計算部と、

を備えたことを特徴とする文書匿名化装置。(5)

(付記 6)

付記 4 又は 5 記載の文書匿名化装置に於いて、前記特定度算出部は、既存文書に基づいて作成した匿名対象表記、人名か周辺表記かの種別及び特定度を組にした特定度データを登録した基準特定度データベースを備え、前記入力文書から抽出した匿名化表記の計算により求めた特定度を、前記基準匿名度データベースに登録している特定度との重み平均をとって正規化することを特徴とする文書匿名化装置。（6）

（付記 7）

付記 4 又は 5 記載の文書匿名化装置に於いて、前記特定度算出部は、文書データベースの既存文書から文書毎に人名や周辺表記を抽出して特定度を計算し、匿名対象表記、人名又は周辺表記の種類及び特定度の組にした特定度データを登録した前記基準特定度データベースを作成するデータベース作成部を備えたことをことを特徴とする文書匿名化装置。（7）

（付記 8）

付記 1 記載の文書匿名化装置に於いて、更に、前記匿名化処理部で使用する閾値を設定変更する匿名化指示部を設けたことを特徴とする文書匿名化装置。（8）

（付記 9）

付記 1 記載の文書匿名化装置に於いて、前記匿名化指示部は、処理文書毎に匿名化処理に使用した閾値を閾値データベースに保存し、新たな入力文書の匿名化処理の際に直前の閾値をデフォルトとして設定することを特徴とする文書匿名化装置。（9）

（付記 1 0）

付記 1 記載の文書匿名化装置に於いて、前記匿名化処理部は、匿名化不要表記を登録して匿名化不要データベースを持ち、入力文書から抽出された匿名化表記の内、前記匿名化不要データベースに登録されている表記は匿名化しないことを特徴とする文書匿名化装置。（1 0）

（付記 1 1）

付記 1 記載の文書匿名化装置に於いて、前記匿名化処理部は、必ず匿名化する表記を登録した匿名化データベースを持ち、該匿名化データベースに登録されている入力文書中の表記は全て匿名化することを特徴とする文書匿名化装置。（1 1）

)

(付記 1 2)

付記 1 記載の文書匿名化装置に於いて、前記匿名化処理部は、入力文書から抽出された匿名化対象表記を伏せ字にすることを特徴とする文書匿名化装置。(1 2)

(付記 1 3)

付記 1 記載の文書匿名化装置に於いて、前記匿名化処理部は、入力文書から抽出された匿名化対象表記を、個人を特定しない一般化された表記に置き換えることを特徴とする文書匿名化装置。(1 3)

(付記 1 4)

付記 1 記載の文書匿名化装置に於いて、前記匿名化処理部は、入力文書から抽出された匿名化対象表記を、該匿名化対象表記の匿名化に使用する閾値以下の特定度を持つ低特定度表記で置き換えることを特徴とする文書匿名化装置。(1 4)

(付記 1 5)

付記 1 記載の文書匿名化装置に於いて、前記匿名化処理部は、入力文書から抽出された匿名化対象表記を暗号化することで匿名化することを特徴とする文書匿名化装置。(1 5)

(付記 1 6)

付記 1 5 記載の文書匿名化装置に於いて、更に、前記匿名化処理部により暗号化により匿名化された匿名化文書を閲覧する際に、暗号化された匿名化表記を復号化して表示させる復号化指示部を設けたことを特徴とする文書匿名化装置。(1 6)

(付記 1 7)

文書を入力する文書入力ステップと、
前記入力文書から匿名対象表記を抽出し、抽出した匿名対象表記がどの程度の強さで個人を特定できるかを評価する特定度を算出する特定度計算ステップと、
所定の閾値より大きい特定度を持つ前記入力文書中の表記を匿名化する匿名化処理ステップと、
を備えたことを特徴とする文書匿名化方法。(1 7)

(付記 1 8)

付記 1 7 記載の文書匿名化方法に於いて、

前記特定度計算ステップは、前記入力文書から人名を抽出し、抽出した人名がどの程度の強さで個人を特定できるかを評価する特定度を算出し、

前記匿名化処理ステップは、所定の閾値よりも大きい特定度をもつ人名を匿名化することを特徴とする文書匿名化方法。

【 0 1 1 2 】

(付記 1 9)

付記 1 7 記載の文書匿名化方法に於いて、

前記特定度計算ステップは、前記入力文書から人名の周辺表記抽出し、抽出された周辺表記がどの程度の強さで個人を特定できるかを評価する特定度を算出し、

前記匿名化処理ステップは、所定の閾値よりも大きい特定度をもつ周辺表記を匿名化することを特徴とする文書匿名化方法。

【 0 1 1 3 】

(付記 2 0)

付記 1 7 記載の文書匿名化方法に於いて、前記特定度算出ステップは、

入力文書から文を切出す文切出しステップと、

切出した文を品詞毎に分解する品詞解析ステップと、

前記品詞解析結果から人名抽出ルールに基づいて人名を抽出する人名抽出ステップと、

統計情報に基づいて抽出した人名の特定度を計算する人名特定度計算ステップと、

を備えたことを特徴とする文書匿名化方法。

【 0 1 1 4 】

(付記 2 1)

付記 1 7 記載の文書匿名化方法に於いて、前記特定度算出ステップは、更に、前記品詞解析結果から構文解析ルールに基づいて文節間の係り受け関係を示す構文木を作成する構文解析ステップと。

前記構文解析ステップで得られた構文木に対し個人特定木抽出ルールに基づいて

個人特定木を個人周辺表記として抽出する個人特定木抽出ステップと、統計情報に基づいて抽出した個人特定木の特定度を計算する木構造特定度計算ステップと、
を備えたことを特徴とする文書匿名化方法。

【0115】

(付記22)

付記20又は21記載の文書匿名化方法に於いて、前記特定度算出ステップは、前記入力文書から抽出した匿名化表記の計算により求めた特定度を、既存文書に基づいて作成した匿名対象表記、人名か周辺表記かの種別及び特定度を組にした特定度データを、登録した基準特定度データベースに登録している特定度との麻績み平均をとって正規化することを特徴とする文書匿名化方法。

【0116】

(付記23)

付記20又は21記載の文書匿名化方法に於いて、前記特定度算出ステップは、文書データベースの既存文書から文書毎に人名や周辺表記を抽出して特定度を計算し、匿名対象表記、人名又は周辺表記の種類及び特定度の組にした特定度データを登録した前記基準特定度データベースを作成するデータベース作成ステップを備えたことを特徴とする文書匿名化方法。

【0117】

(付記24)

付記17記載の文書匿名化方法に於いて、更に、前記匿名化処理ステップで使用する閾値を設定変更する匿名化指示ステップを設けたことを特徴とする文書匿名化方法。

【0118】

(付記25)

付記17記載の文書匿名化方法に於いて、前記匿名化指示ステップは、処理文書毎に匿名化処理に使用した閾値を閾値データベースに保存し、新たな入力文書の匿名化処理の際に直前の閾値をデフォルトとして設定することを特徴とする文書匿名化方法。

【 0 1 1 9 】

(付記 2 6)

付記 1 7 記載の文書匿名化方法に於いて、前記匿名化処理ステップは、入力文書から抽出された匿名化表記の内、匿名化不要データベースを参照して登録されている表記は匿名化しないことを特徴とする文書匿名化方法。

【 0 1 2 0 】

(付記 2 7)

付記 1 7 記載の文書匿名化方法に於いて、前記匿名化処理ステップは、必ず匿名化する表記を登録した匿名化データベースに登録されている入力文書中の表記は全て匿名化することを特徴とする文書匿名化方法。

【 0 1 2 1 】

(付記 2 8)

付記 1 7 記載の文書匿名化方法に於いて、前記匿名化処理ステップは、入力文書から抽出された匿名化対象表記を伏せ字にすることを特徴とする文書匿名化方法。

【 0 1 2 2 】

(付記 2 9)

付記 1 7 記載の文書匿名化方法に於いて、前記匿名化処理ステップは、入力文書から抽出された匿名化対象表記を、個人を特定しない一般化された表記に置き換えることを特徴とする文書匿名化方法。

【 0 1 2 3 】

(付記 3 0)

付記 1 7 記載の文書匿名化方法に於いて、前記匿名化処理ステップは、入力文書から抽出された匿名化対象表記を、該匿名化対象表記の匿名化に使用する閾値以下の特定度を持つ低特定度表記で置き換えることを特徴とする文書匿名化方法。

【 0 1 2 4 】

(付記 3 1)

付記 1 7 記載の文書匿名化方法に於いて、前記匿名化処理ステップは、入力文書から抽出された匿名化対象表記を暗号化することで匿名化することを特徴とする

文書匿名化方法。

【 0 1 2 5】

(付記 3 2)

付記 3 2 記載の文書匿名化方法に於いて、更に、前記匿名化処理ステップにより暗号化により匿名化された匿名化文書を閲覧する際に、暗号化された匿名化表記を復号化して表示させる復号化指示ステップを設けたことを特徴とする文書匿名化方法。

【 0 1 2 6】

(付記 3 3)

コンピュータに、
文書を入力する文書入力ステップと、
前記入力文書から匿名対象表記を抽出し、抽出した匿名対象表記がどの程度の強さで個人を特定できるかを評価する特定度を算出する特定度計算ステップと、
所定の閾値より大きい特定度を持つ前記入力文書中の表記を匿名化する匿名化処理ステップと、
を実行させるための匿名化プログラムを記録したコンピュータ読取り可能な記録媒体。(18)

【 0 1 2 7】

【発明の効果】

以上説明してきたように本発明によれば、文書中の個人を特定するような表現に対し、それがどの程度の強さで個人を特定できるのかを、匿名化を行う前に評価しつつ、要求される匿名化の水準（閾値）に応じて対象となる表記を匿名化して適切に隠蔽化でき、これによって文書を必要な度合いで自動ないし半自動で匿名化でき、匿名化作業を効率化し、作業コストを大幅に低減することができる。

【図面の簡単な説明】

【図 1】

本発明の概略説明図

【図 2】

本発明の機能構成のブロック図

【図 3】

本発明による文書匿名化処理のフローチャート

【図 4】

図 2 の特定度計算部の機能構成のブロック図

【図 5】

構文解析で得られた周辺表記の構文木の説明図

【図 6】

図 5 の構文木から抽出された個人特定木の説明図

【図 7】

図 4 の特定度計算処理のフローチャート

【図 8】

図 2 の特定度計算部に設けているデータベース作成部の機能構成のブロック図

【図 9】

基準特定度データベースの説明図

【図 1 0】

図 8 の基準特定度データベース作成処理のフローチャート

【図 1 1】

本発明における匿名化処理のフローチャート

【図 1 2】

図 2 の閾値データベースの説明図

【図 1 3】

図 1 1 の置換処理のフローチャート

【図 1 4】

本発明で処理する原文作業画面の説明図

【図 1 5】

低レベルの閾値を指示した場合の本発明による匿名化文書の画面説明図

【図 1 6】

高レベルの閾値を指示した場合の本発明による匿名化文書の画面説明図

【符号の説明】

- 1 0 : 文書入力部
- 1 2 : 特定度計算部
- 1 4 : 基準特定度データベース
- 1 5 : データベース作成部
- 1 8 : 匿名化処理部
- 2 0 : 匿名化指示部
- 2 2 : 匿名化不要データベース
- 2 4 : 匿名化データベース
- 2 6 : 閾値データベース
- 2 8 : 作業表示部
- 3 0 : 匿名化文書記憶部
- 3 2 : 復号化指示部
- 3 4 : 判定部
- 3 6 : 閲覧データ作成部
- 3 8 : 閲覧表示部
- 4 0 : 文切出部
- 4 2 : 品詞解析部
- 4 4 : 人名処理部
- 4 4 - 1 : 人名処理部（データベース作成部内）
- 4 6 : 周辺表記処理部
- 4 6 - 1 : 周辺表記処理部（データベース作成部内）
- 4 8 : 人名抽出部
- 5 0 : 人名抽出ルール
- 5 2 : 人名特定度計算部
- 5 4 : 構文解析部
- 5 6 : 個人特定木抽出部（周辺表記抽出部）

5 8 : 木構造特定度計算部 (周辺表記特定度計算部)

6 0 : 構文解析ルール

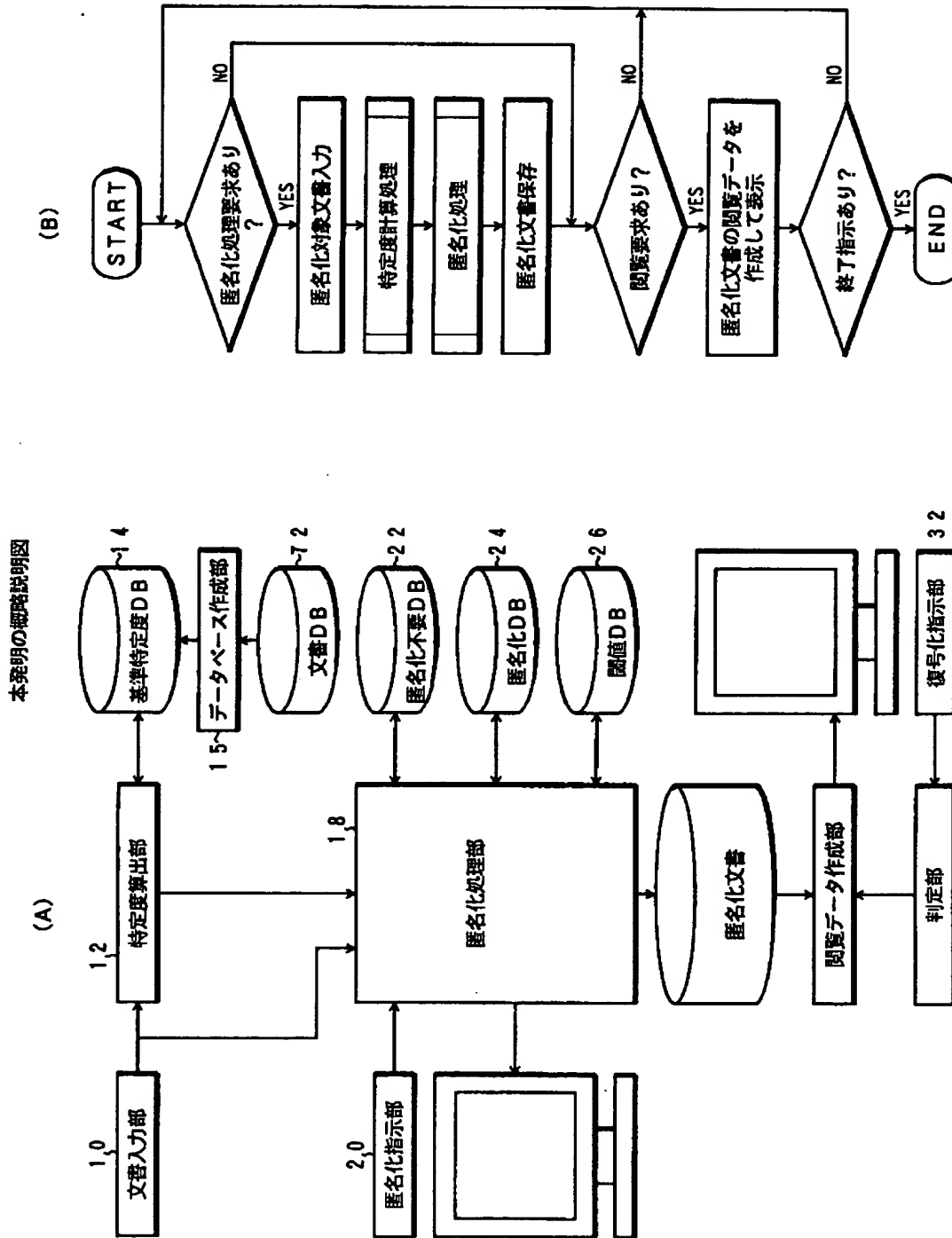
6 2 , 8 6 : 個人特定木抽出ルール

7 2 : 文書データベース

7 4 : 文書切出部

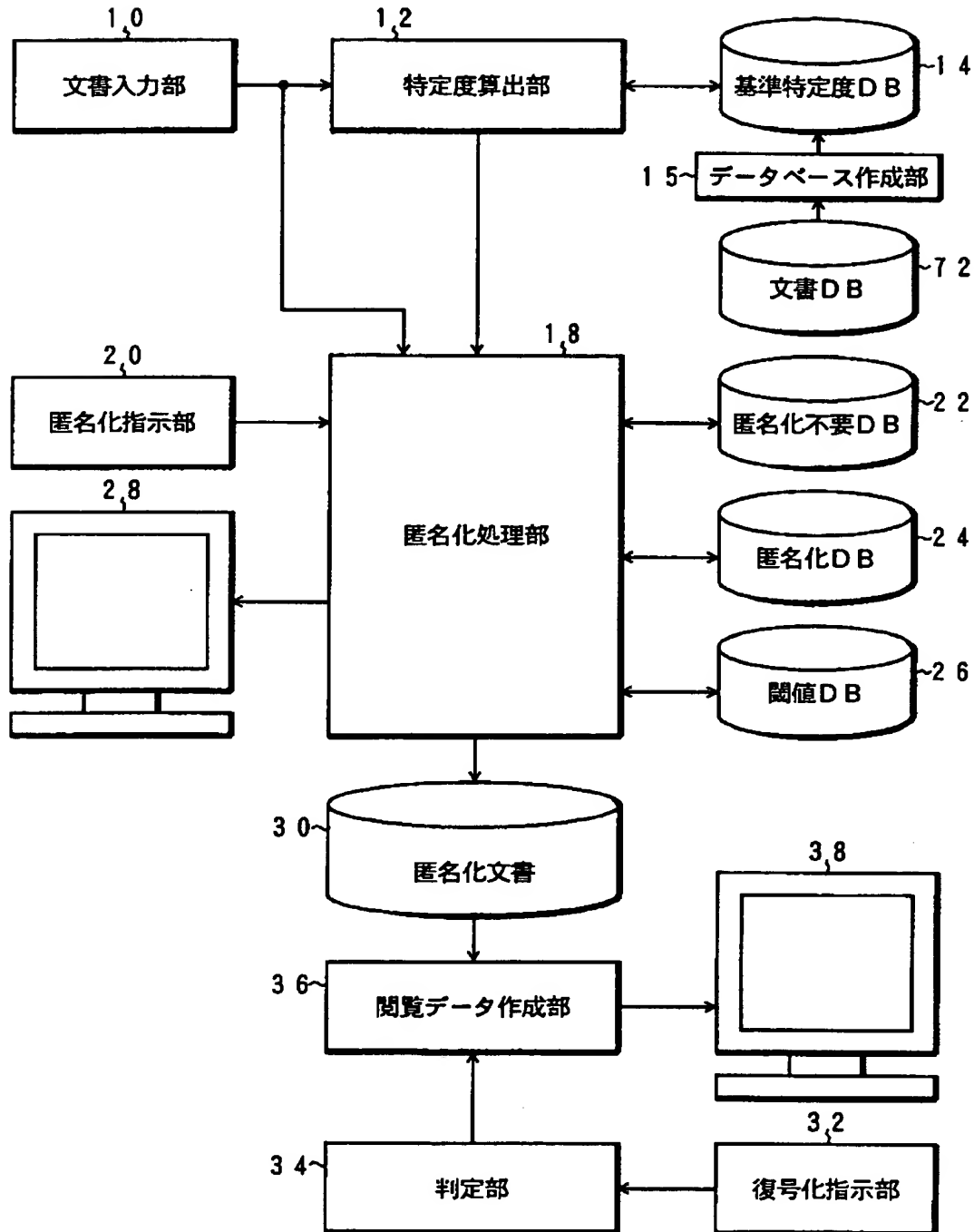
【書類名】 図面

【図 1】



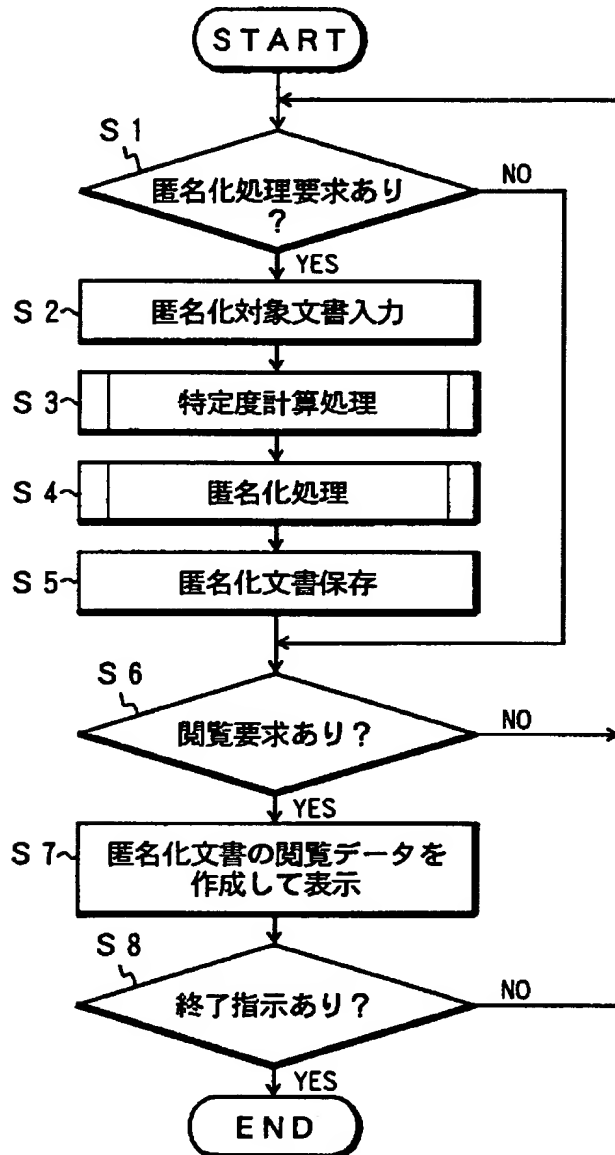
【図 2】

本発明の機能構成のブロック図



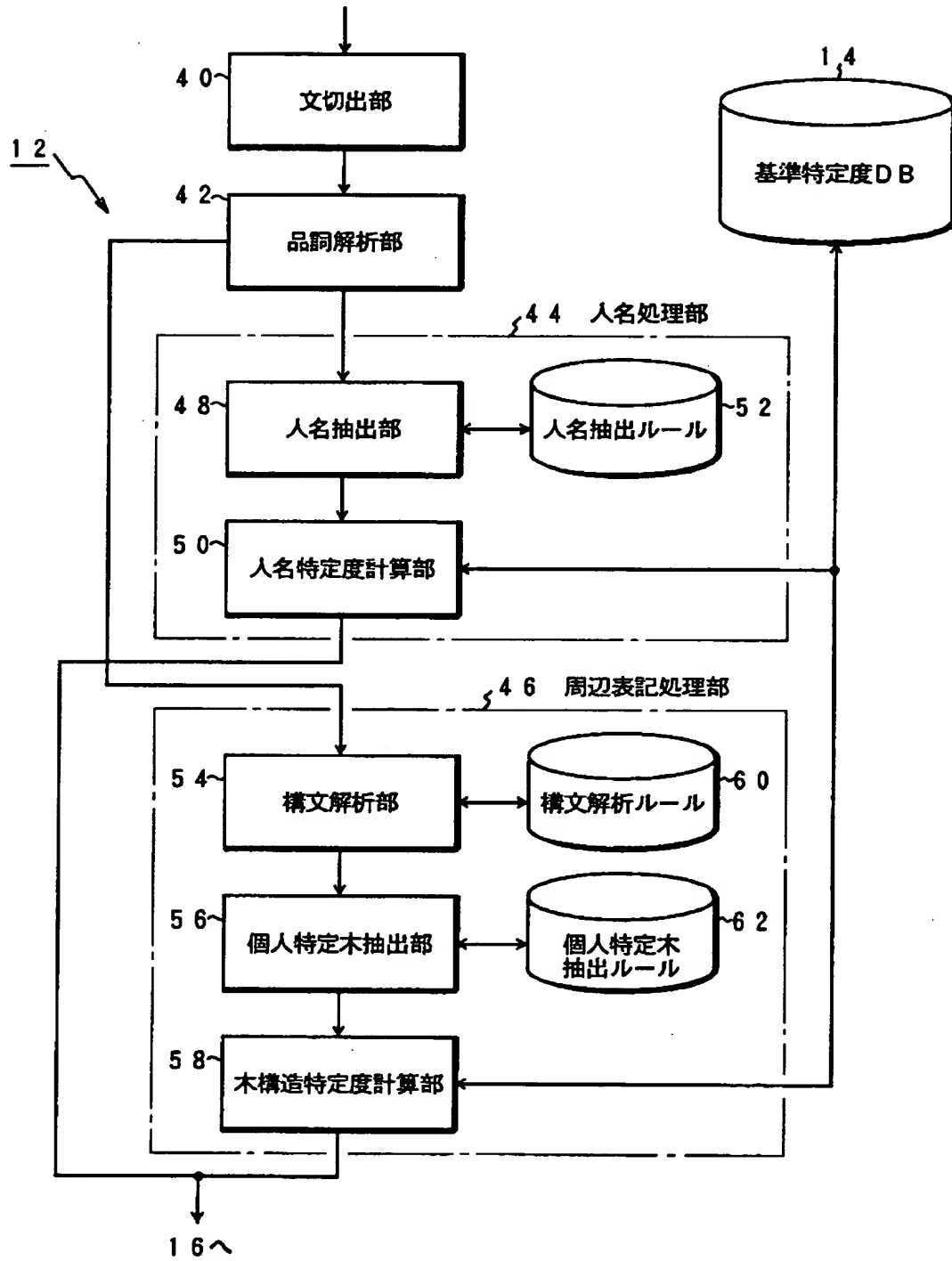
【図 3】

本発明による文書匿名化処理のフローチャート



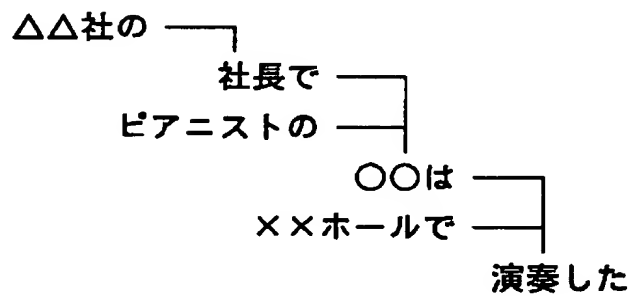
【図 4】

図 2 の特定度計算部の機能構成のブロック図



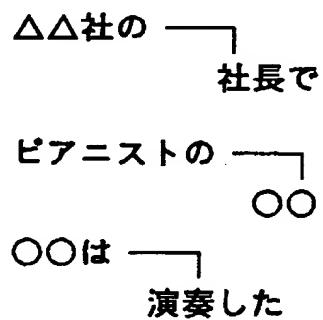
【図 5】

構文解析で得られた周辺表記の構文木の説明図



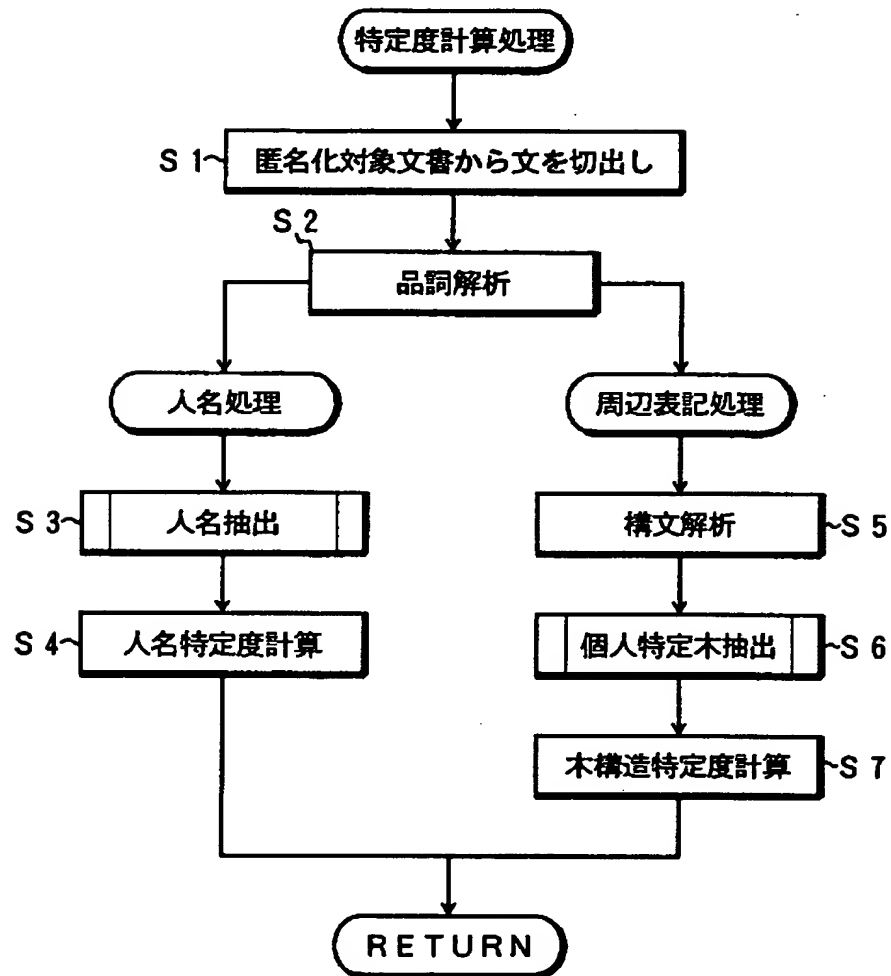
【図 6】

図 5 の構文木から抽出された個人特定木の説明図



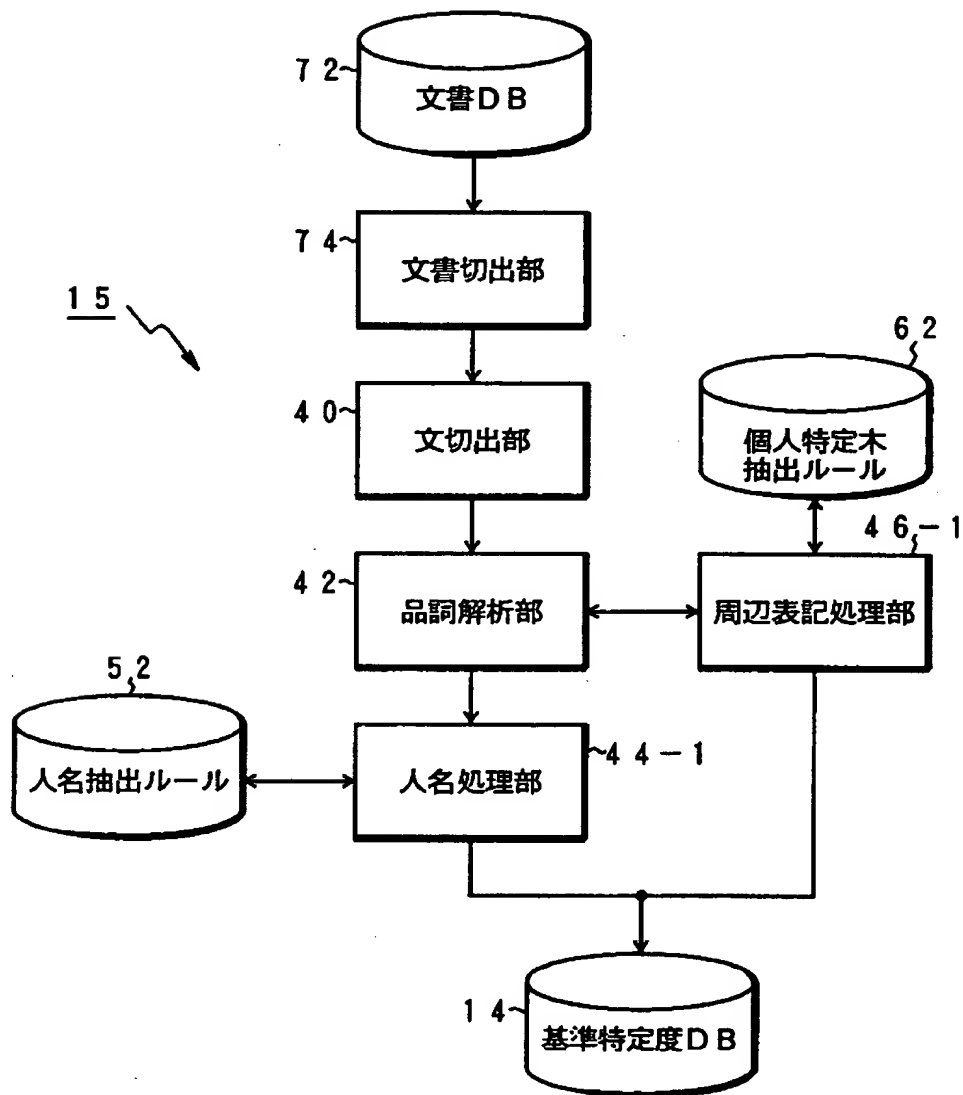
【図 7】

図 4 の特定度計算処理のフローチャート



【図 8】

図 2 の特定度計算部に設けているデータベース作成部の機能構成のブロック図



【図 9】

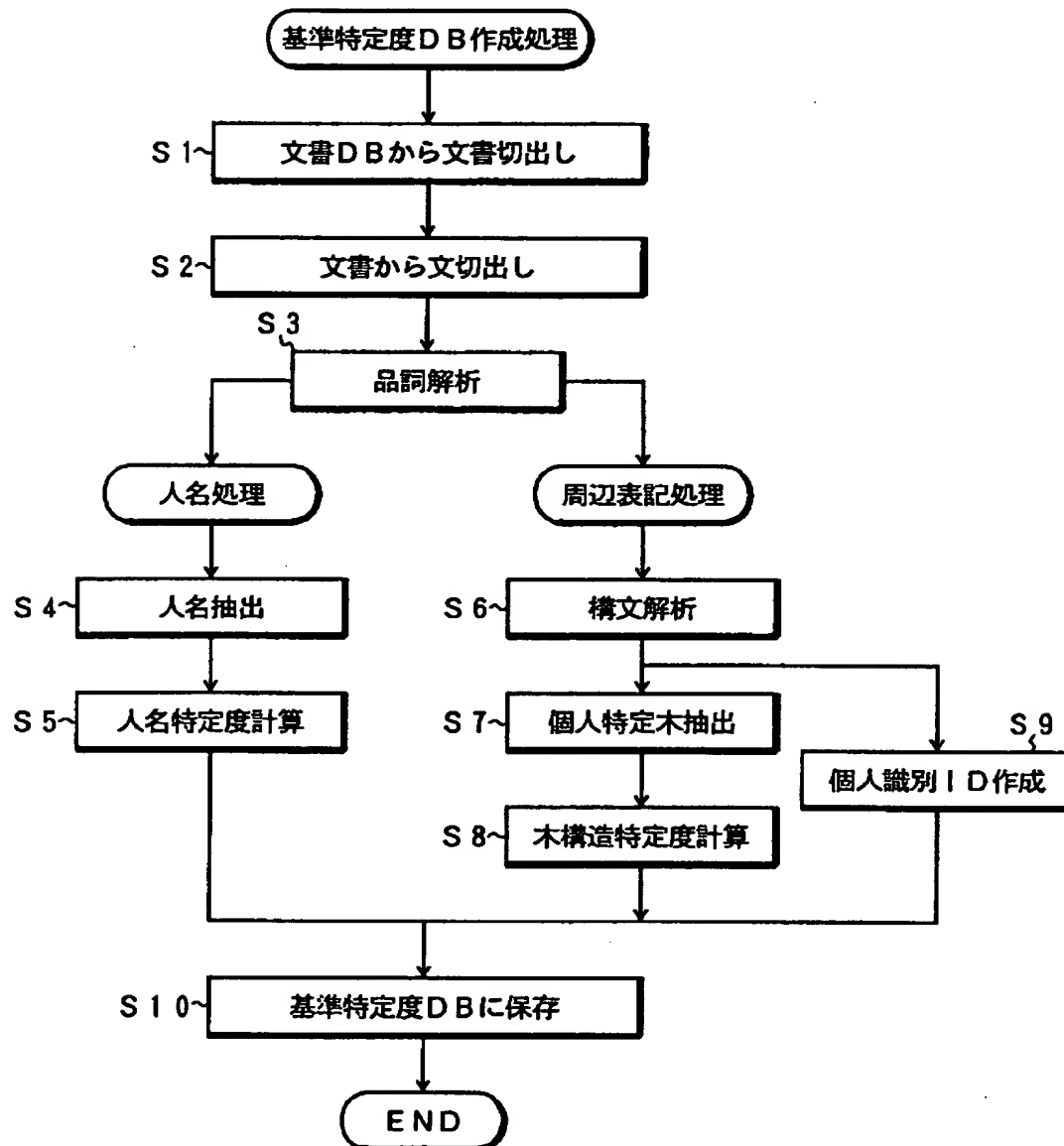
図 2 の閾値データベースの説明図

14

個人識別 I D	表記の種類	表記	確率
P 0 0 1	人名	松岡	0.3
P 0 0 1	人名	松岡英達	0.9
P 0 0 1	周辺	h-matsu@karino.kaisha.co.jp	1.0
P 0 0 1	周辺	情報機器営業部の	0.2
....
P 0 0 3	人名	松岡	0.2
P 0 0 3	周辺	埼玉県草加市	0.6
....
P 0 0 9	人名	埼玉県草加市	0.6
....

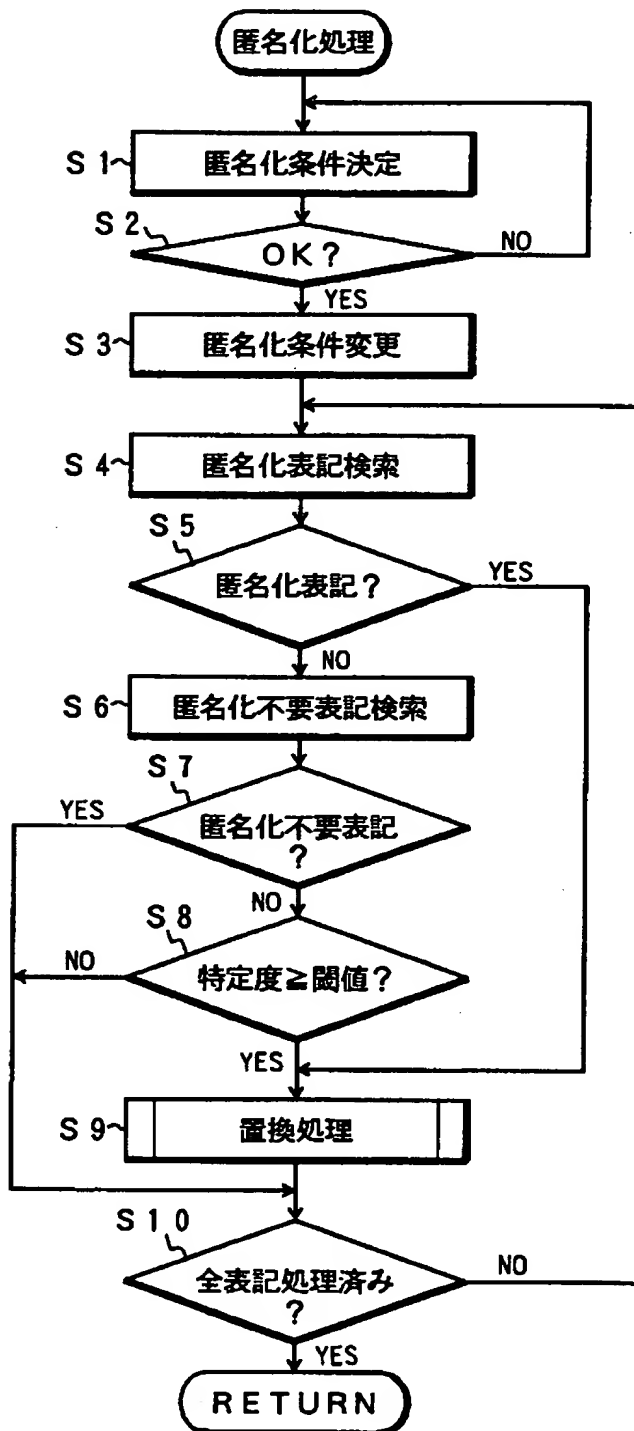
【図 1 0】

図 8 の基準特定度データベース作成処理のフローチャート



【図 1 1】

本発明における匿名化処理のフローチャート



【図 1 2】

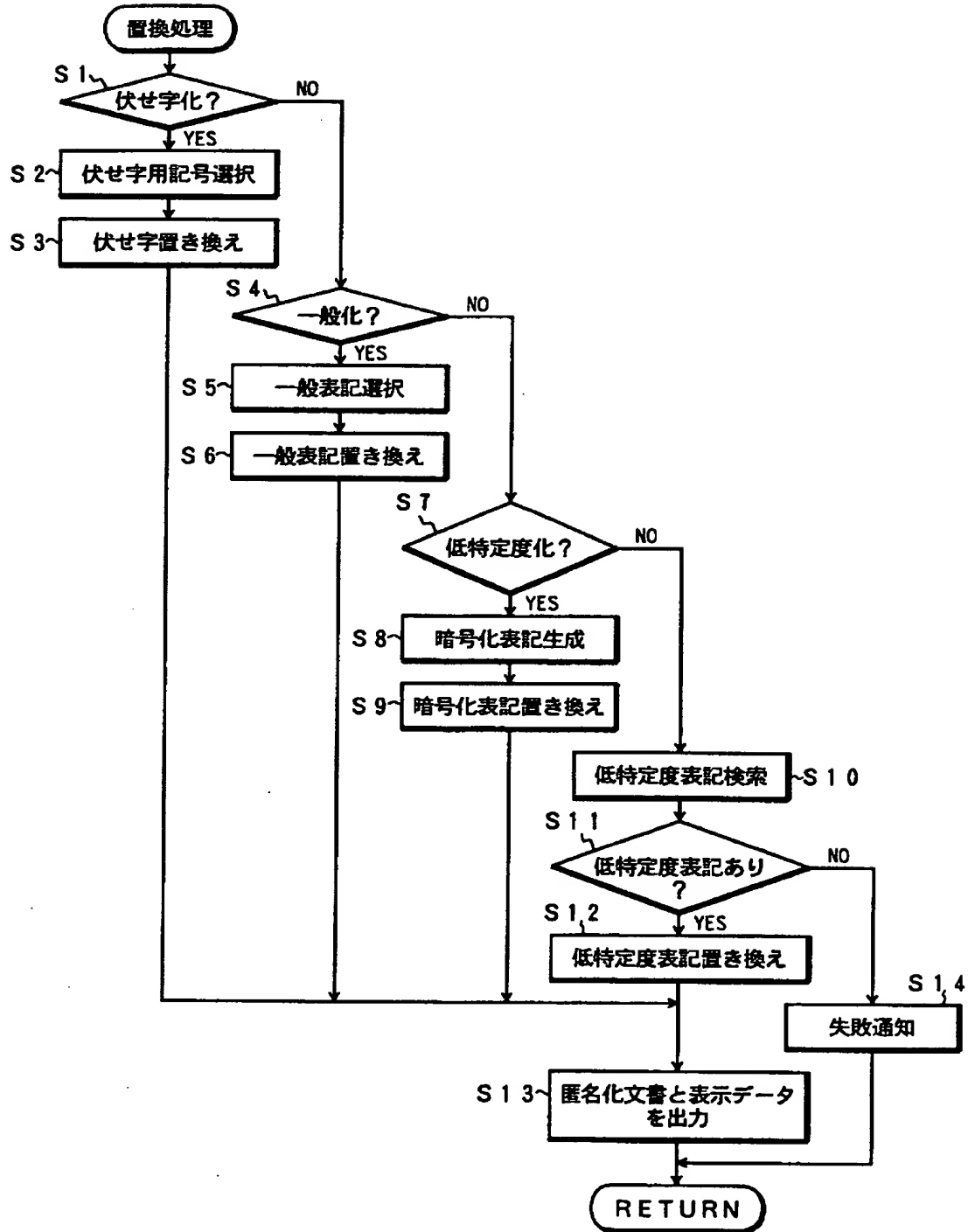
図 2 の閾値データベースの説明図

2 6

処理文書分類	閾 値	匿名化方法
0 0	直前閾値	伏せ字化
0 1	T H 1	一般化
0 2	T H 2	一般化
0 3	T H 3	低特定度化
0 4	T H 4	暗号化
⋮	⋮	⋮
n	T H n	伏せ字化

【図 13】

図 11 の置換処理のフローチャート



【図 14】

本発明で処理する原文作業画面の説明図

88

90

92

From: h-matsuoka@karino.kaisha.co.jp
 To: satou@mo.tantou.co.jp
 Date: Mon, 23 Aug 1999 10:22:34 +0900
 Subject: ご挨拶

佐藤様

情報媒体、情報機器営業部の松岡と申します。

メールで大変失礼致します。
 御社ホームページでセキュリティ機能付きMO装置の記事
 を拝見しました。
 私も弊社のMO装置の販売に関係しておりますので、一度
 お伺いさせて頂き、ご挨拶させて頂けませんかでしょうか。

気楽に情報交換をさせて頂ければ幸甚に存じます。

佐藤様の都合の良い日、時間、場所をお知らせ頂ければ幸いです。
 お忙しい中恐縮ですがよろしくお願い申し上げます。

99年8月23日

情報媒体株式会社
 情報機器営業部
 松岡英達
 h-matsuoka@karino.kaisha.co.jp

TEL 044-754-2671 FAX 044-754-2570
 神奈川県川崎市中原区上小田中4-1-1

レベル

原文 ▼

92-1

匿名化

実行

94

【図 15】

低レベルの閾値を指示した場合の本発明による匿名化文書の画面説明図

8 8

9 6

9 2

From: xxxxx@xxxxx.co.jp
 To: yyyyy@yyyyy.co.jp
 Date: Mon, 23 Aug 1999 10:22:34 +0900
 Subject: ご挨拶

佐藤様

〇〇〇〇、××××営業部の松岡と申します。

メールで大変失礼致します。
 御社ホームページでセキュリティ機能付きMO装置の記事
 を拝見しました。
 私も弊社のMO装置の販売に関係しておりますので、一度
 お伺いさせて頂き、ご挨拶させて頂けませんかでしょうか。

気楽に情報交換をさせて頂ければ幸甚に存じます。

佐藤様の都合の良い日、時間、場所をお知らせ頂ければ幸いです。
 お忙しい中恐縮ですがよろしくお願い申し上げます。

9 9 年 8 月 2 3 日

〇〇〇〇株式会社
 ××××営業部
 松岡△△
 yyyyy@yyyyy.co.jp

TEL ZZZ-ZZZ-ZZZZ FAX WWW-WWW-WWW
 □□□□□□□□□□□□□□

レベル

低
▼

9 2 - 2

匿名化

実行

9 4

【図 16】

高レベルの閾値を指示した場合の本発明による匿名化文書の画面説明図

88

96

From: xxxxx@xxxxx.co.jp
 To: yyyyy@yyyyy.co.jp
 Date: Mon, 23 Aug 1999 10:22:34 +0900
 Subject: ご挨拶

▽▽様

〇〇〇〇、××××営業部の△△と申します。

メールで大変失礼致します。
 御社ホームページでセキュリティ機能付きMO装置の記事
 を拝見しました。
 私も××××××××××××××××××××××××、一度
 お伺いさせて頂き、ご挨拶させて頂けませんかでしょうか。

気楽に情報交換をさせて頂ければ幸甚に存じます。

▽▽様の都合の良い日、時間、場所をお知らせ頂ければ幸いです。
 お忙しい中恐縮ですがよろしくお願い申し上げます。

99年8月23日

〇〇〇〇株式会社
 ××××営業部
 △△△△
 yyyyy@yyyyy.co.jp

TEL ZZZ-ZZZ-ZZZZ FAX WWW-WWW-WWWW
 □□□□□□□□□□□□□□□□

92

レベル

高
▼

92-3

匿名化

実行

94

【書類名】 要約書

【要約】

【課題】 個人情報の隠蔽化を機械化して作業コストを低減し、必要に応じて隠蔽化の度合を調整可能とする。

【解決手段】 文書匿名化装置は文書を入力する文書入力部 1 0 と、入力文書から匿名対象表記を抽出し、抽出した匿名対象表記がどの程度の強さで個人を特定できるかを評価する特定度を算出する特定度計算部 1 2 と、所定の閾値より大きい特定度を持つ入力文書中の表記を匿名化する匿名化处理部 1 8 とを備える。特定度計算部 1 8 は、入力文書から人名と周辺表記を抽出し、抽出した人名と周辺表記がどの程度の強さで個人を特定できるかを評価する特定度を算出し、匿名化处理部 1 8 は、所定の閾値よりも大きい特定度をもつ人名と周辺表記を、伏せ字化、一般化、低特程度化、暗号化等により匿名化する。

【選択図】 図 1

出 願 人 履 歴 情 報

識別番号 [000005223]

1. 変更年月日	1996年 3月26日
[変更理由]	住所変更
住 所	神奈川県川崎市中原区上小田中4丁目1番1号
氏 名	富士通株式会社